Equating Arizona's Instrument to Measure Standards (AIMS) in High School Reading, Writing and Mathematics to Cambridge IGCSE and ACT *QualityCore* English and Mathematics End-of-Course Examinations – Interim Report

Stephen G. Schilling
University of Michigan

Paul G. Perrault
University of Michigan

April 14, 2014

STEPHEN SCHILLING is an assistant research scientist and psychometrician at the Institute for Social Research's Survey Research Center at the University of Michigan, 426 Thompson St., Room 2370 Perry Building, University of Michigan, Ann Arbor, MI 48134; e-mail: *schillsg@umich.edu.* His research interests include multidimensional IRT models, Markov Chain Monte Carlo Methods, and applications of educational assessments.

PAUL G. PERRAULT is a research investigator at the Institute for Social Research's Survey Research Center at the University of Michigan, 426 Thompson St., Room 2362 Perry Building, University of Michigan, Ann Arbor, MI 48134; e-mail: *paulperr@umich.edu.* His research examines the effects of educational interventions and education policy on student achievement and teachers' perceptions of instructional change.

Work on this study is supported by a grant from the National Center on Education and the Economy. The authors would like to thank Edward Haertel, Emeritus Professor, Stanford University, for his valuable comments on the proposed work and the Center for the Future of Arizona for their technical support to the state's *Move on When Ready* initiative that framed this effort.

Introduction

This paper was developed to aid in the establishment of equivalent proficiency scores for Arizona's Instrument to Measure Standards (AIMS) examinations and the end-of-course assessments utilized in the *Move On When Ready/Excellence for All* (*MOWR/E4A*) initiative, as for a variety of education policy purposes it is useful to be able to translate how scores on one set of exams map onto scores on the other. This paper thus focuses on the assessments currently approved for use in *MOWR/E4A* by the Arizona State Board of Education, ACT's *QualityCore* and the University of Cambridge's International General Certificate of Secondary Education (IGCSE)

Methods and Approaches for Establishing Equivalent Proficiency

To address this requirement the paper first spells out the methodological challenges associated with these kinds of technical questions and then offers an approach to address them. Next it takes the methods presented and applies them to establish equivalent proficiency scores for the AIMS examinations and the *MOWR/E4A* aligned instructional provider end-of-course assessments to help establish relevant benchmarks for performance on the *MOWR/E4A* examinations.

Equivalent Proficiency – Linking

Linking refers to established methods of connecting scores on different tests and reporting tests scores on a common scale (Feuer, Holland, Green, Bertenthal, & Cadell, 1999). The most useful analogy for linking is converting temperature back and forth from the Fahrenheit scale to the Celsius scale. That relationship is expressed as:

$$°\mathrm{F} = \left({}^{9}\!/_{5}\right)°\mathrm{C} + 32 \quad \text{or} \quad ℃ = {}^{5}\!/_{9}\,(°\mathrm{F} - 32).$$

Using that linking relationship we can interpret 35 ℃ in London as 95 °F and so recognize that 35 ℃ in London is a hot summer day. This is an example of a linear linking relationship. In order to get the formula for ℃ we just use algebra to invert the formula for °F. Thus we say that the relationship is invertible. The analogy is also useful in that we recognize and accept that both °F and ℃ refer to the same content – the position of mercury in a thermometer. And the linking formula makes us indifferent to whether temperature is reported on the Fahrenheit or Celsius; if we are given a temperature in ℃ we can use the linking formula to convert to °F and easily interpret the results.

The temperature analogy is useful as an ideal that we aim for in linking tests. We would like to be able to accept the content of the two tests as equal, have an invertible formula for linking the two tests, and be indifferent to whether scores are reported on text X or test Y because we have a formula for converting from one test score to another. However, linking tests is not as simple as in the temperature example. Because of possible differences in content, populations being assessed, and circumstances of testing, each linking endeavor must be assessed on a case-by-case basis to determine the strength of the inferences that can be provided. Feuer et al. (1999) classify linking methods according to the strength of the inferences they provide.

The strongest linking relationship that can be established between two tests is equating. If two tests are equated then you are indifferent to whether one test or the other is taken, and benchmarks on either test are interchangeable. The generally accepted requirements for strict equating of tests are given below.

Requirements for Equating

i. Equal Construct
ii. Equal Reliability
iii. Symmetry
iv. Equity – Indifference to use of either test
v. Population Invariance

Equating is most frequently employed when comparing different forms of same test. Here the content of the different forms is the same by design. The circumstances surrounding the test administration such as high stakes vs. low stakes, formative assessment vs. summative assessment, end of year vs. beginning of year, etc. are the same. The test items have the same format and the different forms are constructed to have equal reliability. The relationship between scores on the two tests is the same for each possible subpopulation of intended examinees, so we have population invariance.

Symmetry is a function of the technical methods of test linking employed, such as linear equating or equipercentile equating. Symmetry is important because it allows us to invert the linking relationship. If X refers to the test to be linked and Y refers to the target test for linking (e.g., Arizona AIMS High School (HS) Mathematics test), our objective can be either to determine the value of Y corresponding to a specific value of X or the value of X corresponding to a particular value of Y. In the first case we are usually interested in providing students a criterion test score on the target test, while in the second case we often desire the value of X corresponding to competency thresholds on the second. When we can easily invert the functional linking relationship for producing a Y score for each value of X we can easily meet both objectives. This means the linking relationship is “invertible” and symmetry is preserved. Not all possible relationships are invertible. For example the regression of X on Y is not the inverse of the regression of Y on X. If we have invertible linking between different forms, we can use the invertible scores interchangeably.

If the different forms are then used for the same purposes, such as evaluating year-end performance we should be indifferent to the use of different forms – test equity. Then all five conditions for equating are met and we can confidently make the same inferences for alternative forms of the same test. The closer we are to meeting all five conditions the more confident we are with regards to our inferences. When we go beyond linking of alternative forms of the same test we need to closely consider each of these conditions.

Our objective in this report is to establish a linking relationship between the Arizona AIMS HS Reading, Writing and Mathematics examinations with their corresponding Cambridge IGCSE and ACT *QualityCore* end-of-course exams. The objectives of the exams are the same – measuring proficiency in the respective subjects. The populations are the same: 10th grade Arizona high school students[1]

The Cambridge IGCSE English First Language assessment consists of numerically scored written papers and a coursework portfolio.[2] The content domains for the Cambridge IGCSE English First Language exam includes Reading and Writing with an optional and separately scored domain on Speaking and Listening. The numerically scored written papers account for 50% of a student's score (80% testing the Reading domain and 20% the Writing domain), while the coursework portfolio accounts for the other 50% of a student's score (20% testing Reading and 80% Writing). The ACT *QualityCore* English 10 consists of two equally weighted sections. The first consists of approximately 40 multiple choice questions, while the second calls for an essay in response to a text. In contrast, the AIMS HS Reading exam consists of around 30 multiple choice reading comprehension items. The AIMS HS Writing test consists of a smaller number of multiple choice questions assessing student's knowledge of the practice of Writing, an additional multiple choice section on writing "Think Throughs," and an extended writing response.

The Cambridge IGCSE Mathematics assessment consists of approximately 40 short answer questions and a smaller number of structured questions. These components are weighted at 35% and 65% respectively. Both ACT *QualityCore* Algebra I & Geometry consist of two equally weighted sections. The first consists of about 40 multiple choice questions, while the second is organized around constructed response type questions. In contrast the AIMS HS Mathematic test consists of around 40 multiple choice questions, a short "Think Throughs" and practice application section.

The different structure of the Cambridge IGCSE, the ACT *QualityCore*, and Arizona AIMS exams, especially for the English/Reading/Writing exams presents challenges for linking in the sense that our confidence in our inferences is not as strong as linking alternative forms of the same test. However, linking the two tests in this situation is still valuable. Feuer, Holland, Green, Bertenthal, & Cadell (1999) note that when tests are built to different content frameworks or test specifications the strict conditions for equating will almost always be violated to some degree. "When the scores from two different tests are put on the same scale the results are said

[1] Students who do not meet proficiency in any AIMS subject are required to retake examinations in order to graduate. Therefore, we used only 10th grade student AIMS data for first time takers.

[2] Cambridge IGCSE examinations are scored around Percentage Uniform Marks from 20 – 100; ACT *QualityCore* reports their data by scale scores from 125-175.

to be comparable or calibrated" (Feuer, Holland, Green, Bertenthal, & Cadell, 1999, page 18-19). The statistical methods used in equating can be used to achieve comparable scores but more caution needs to be used in their interpretation. Additional information and judgment needs to be used to be confident that the inferences from those comparable scores are reasonable.

The linking methods we use can help strengthen our ability to make inferences based on the linked scores. Linking based on equivalent groups reduces population differences. For Arizona we have a relatively large number of students taking the AIMS exams and either the Cambridge IGCSE or ACT *QualityCore* exams.

In what follows we first describe the linking methods we used. Then we provide results for equating the Cambridge IGCSE Mathematics end-of-course exam to the AIMS HS Math exam. We then provide results for equating Cambridge IGCSE English First Language end of course exams with both the AIMS HS Reading and Writing exams. Then we provide results for equating ACT *QualityCore* Algebra I and Geometry exams to the AIMS HS Mathematics exam. Finally we provide results for equating the ACT *QualityCore* English 10 exam with both the AIMS HS Reading and Writing exams.

Linking Methods

We begin by describing some basic terms and notation. We will be using the term "equating methods" to refer to the methods used to link the tests. As noted above, when the strict conditions of equating are not completely met we can still use equating methods to create comparable scores. In this document we will be referring to observed score equating as opposed to model based IRT equating. X refers to the test to be linked (Cambridge IGCSE English First Language or Mathematics or ACT *QualityCore* English 10 or Algebra I/Geometry). Y refers to the target test for linking (the respective Arizona AIMS HS test). Our objective is to find an equating function e(X) that produces a comparable Y score for each X score, or in layman terms to find a comparable Arizona AIMS HS test score for every Cambridge IGCSE or ACT *QualityCore* score.

The simplest equating method is linear equating in which the observed moments of X and Y are matched, yielding:

$$Lin_Y(x) = \mu_Y + (\sigma_Y/\sigma_X)(x - \mu_X).$$

Linear equating is considered a "strong" method because it makes the strong assumption that the relationship between scores are linear and that the scores are matched with respect to all other higher order moments. These assumptions are unlikely to be met if the two tests differ substantially in difficulty.

Equipercentile equating is a generalization of linear equating that allows for test distributions that deviate from normal bell shaped distributions. It is related to linear equating but adds a nonlinear component to account for deviations from normal distributions. If the two tests have approximately equal difficulty and are approximately normally distributed, any equipercentile method will be equivalent to linear equating. However, most test distributions typically show some departures from normality requiring a nonlinear component in the equating relationship.

The basis for equipercentile equating is the observation that if tests X and Y have cumulative density functions F(X) and G(Y) that are continuous, strictly increasing, and invertible then

$$Equi_Y(x) = G^{-1}(F(x))$$

is an equating function that matches all moments of Y.

There are two difficulties with direct application of the above formula. First, test distributions are typically discrete and not continuous. Second, there can be large sampling variability in the discrete probabilities associated with a single score. Consider 160 subjects on a 40-question test. A histogram of the number right score will exhibit a very spiked pattern because of the large number of categories relative to the number of subjects. The histogram will exhibit large sampling variability with respect to the probability of getting any particular number correct on the test.

Both of these problems are addressed via smoothing. The first step is pre-smoothing the discrete distributions via log-linear models (Holland & Thayer, 2000) – usually polynomial log-linear models using powers of the number right score as the polynomial predictor. This serves to smooth the spiked pattern in the histogram of observed number right score. The second step is continuizing the discrete distribution represented by the smoothed histogram. Kernel density estimators of the continuized distribution have recently been introduced (von Davier, Holland, and Thayer, 2004), and can be used to conduct post-smoothing. However, we found that simple linear interpolation was sufficient to achieve appropriate levels of smoothing.

Equipercentile methods with pre- and post-smoothing work well for moderate to large numbers of subjects (100+). In Arizona we had at least 280 examinees for any equating function – for Cambridge IGCSE Mathematics and English First Language exams we had 437 and 879 examinees respectively.

Linking Arizona AIMS HS Mathematics Scale Scores and Cambridge IGCSE Mathematics PUM Scores

In 2012-2013 Arizona had 437 students who took both the AIMS HS Mathematics assessment and the Cambridge IGCSE Mathematics end-of-course examination. These common students formed the basis for common group equipercentile equating. The AIMS Mathematics HS assessment is reported on a scale that ranges from 300 to 700 and has four performance reporting categories corresponding to the NCLB categories of *Minimal, Basic, Proficient* and *Advanced*. The four performance categories are presented in Table 1.

Table 1: Arizona Mathematics Performance Level Descriptors[3]
(From http://www.azed.gov/standards-development-assessment/aims/performance/)

Performance Level	General Descriptor
Exceeds the Standard (537 - 700)	Students who score at this level illustrate a superior academic performance as evidenced by performing substantially beyond the achievement goal for all students. Students who perform at this level demonstrate knowledge, skills, and abilities in fulfillment of the Mathematics Standard. They can create and analyze inductive and deductive arguments and solve problems that contain trigonometric ratios or algebraic concepts.
Meets the Standard (487 - 536)	Students who score at this level demonstrate a solid academic performance on subject matter as reflected by the Mathematics Standard. Students who perform at this level are able to justify the relationships among subsets of the real numbers, solve problems using a system of linear equations, and write the equation of a line. They can calculate surface area and volume of 3-dimensional objects and determine probability in contextual situations. They can solve and factor quadratic equations.
Approaches the Standard (471 - 486)	Students who score at this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show some understanding of the Mathematics Standard's concepts and procedures by being able to solve problems involving similar and congruent figures, organize and display data, and solve and graph linear equations or inequalities. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding
Falls Far Below the Standard (300 - 470)	Students who score at this level may have significant gaps in the knowledge and skills that are necessary to satisfactorily meet the Mathematics Standard. Students will typically require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

The Cambridge IGCSE Mathematics end-of-course examination is reported on a percentage uniform mark (PUM) scale that ranges from 20-99. All Cambridge IGCSE scores have 8 letter reporting categories. The PUM scores associated with each letter grade are given in Table 2.

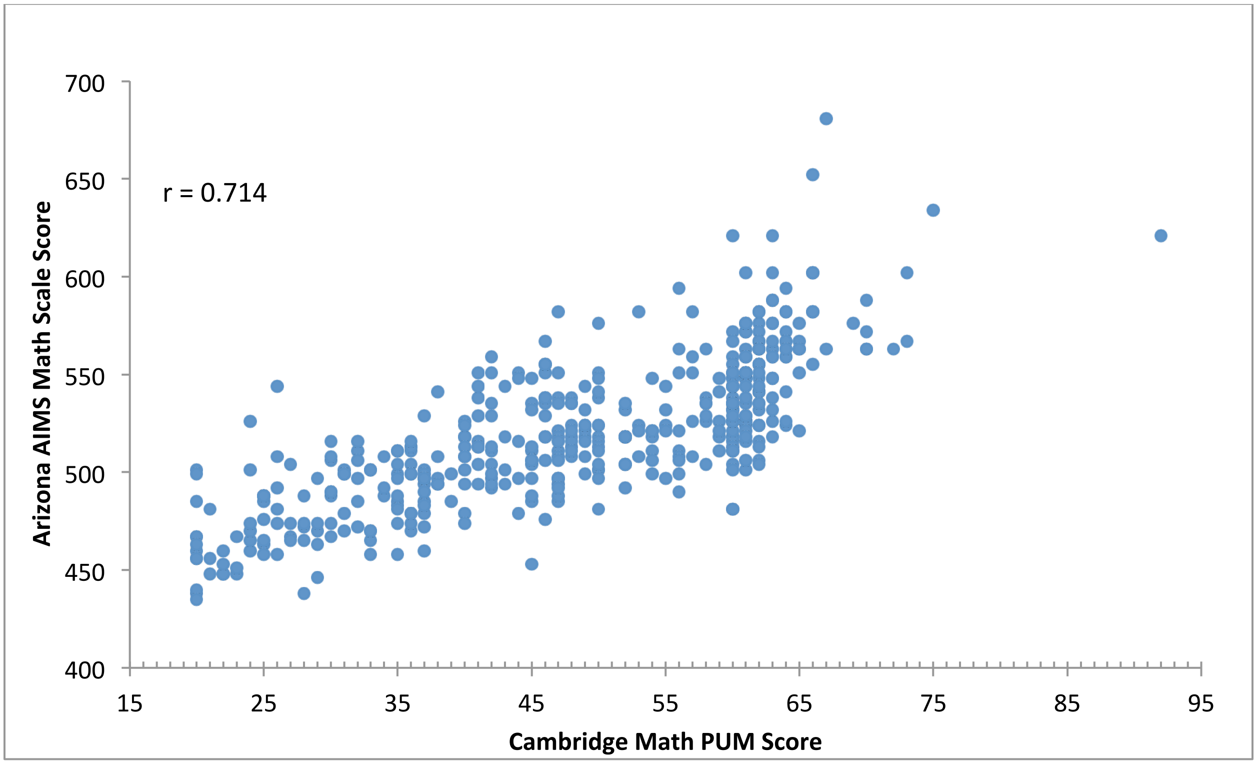
[3] Performance levels can be found at http://www.azed.gov/research-evaluation/files/2012/05/2012aimsscalescores.pdf

Table 2: PUM Scores Associated with Each IGCSE Grade

Grade Threshold	Percentage Uniform Mark
A*	90 – 99
A	80 – 89
B	70 – 79
C	60 – 69
D	50 – 59
E	40 – 49
F	30 – 39
G	20 – 29

A scatter plot of the AIMS HS Mathematics scores versus the Cambridge IGCSE Mathematics scores is given in Figure 1. The two scores show a moderate association, with a correlation of 0.714 (see below).

Figure 1: Scatterplot of AIM HS Mathematics Scale Score vs. Cambridge IGCSE PUM Score



Our first step in this linking was to examine the distributions and smooth the AIMS HS Mathematics and Cambridge IGCSE Mathematics distribution. Pre-smoothing of the Cambridge IGCSE Mathematics and AIMS HS Mathematics distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 3.

Table 3: Fit for Math Pre-smoothing

	AIMS Math	Camb Math
Model	AIC	AIC
Log-lin C=2	273.76	1312.19
Log-lin C=3	272.33	1219.77
Log-lin C=4	272.89	960.31
Log-lin C=5	274.89	947.23
Log-lin C=6	278.99	658.47
Log-lin C=7	278.76	657.51
Log-lin C=8	280.24	560.67

From our analysis we determined that a polynomial log-linear model of order 3 had the best fit for the AIMS Mathematics distribution, while a polynomial of order 8 produced the best fit for the Cambridge IGCSE Mathematics distribution.

Figure 2 presents the equipercentile linking relationship between the Arizona AIMS HS Mathematics examination and the Cambridge IGCSE Mathematics examination. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 2. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Graphically, Figure 2 shows the Cambridge IGCSE Mathematics equivalent of “approaches”, “meets” and “exceeds” standards thresholds for the Arizona AIMS HS Mathematics exam. Table 4 (below) summarizes the reporting thresholds (below), while Table 5 (below) displays the full presentation of Cambridge IGCSE Mathematics equivalent scores.

Evaluating the Cambridge Equivalent Scores

For convenience we indicate by color code in Table 5 where the Cambridge equivalent AIMS score either equals or first exceeds the AIMS Thresholds given in Table 4. For example, the

AIMS "Approaches" threshold in Table 4 is 471. In Table 5 a Cambridge Mathematics score of 24 gives an AIMS equivalent score of 469 while a Cambridge Mathematics score of 25 gives an AIMS equivalent score of 472. Therefore, the Cambridge equivalent "Approaches" threshold is set at 25 because it produces the first AIMS equivalent value that equals or exceeds the "Approaches" threshold. A similar process was used to set the Cambridge equivalent "Exceeds" threshold.

Figure 2: Cambridge IGCSE Mathematics and AIMS HS Math Equipercentile Equating Relationship

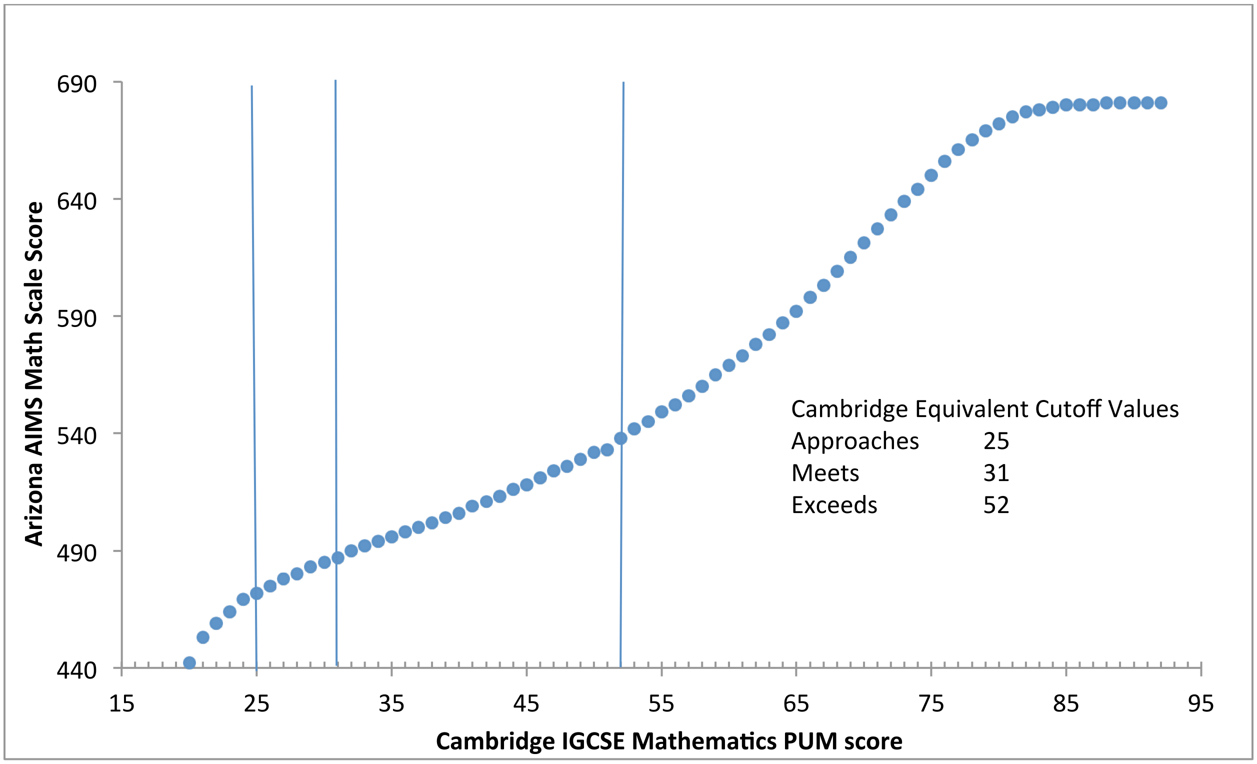


Table 4: Cambridge Equivalent Thresholds

Threshold	AIMS Math	Camb Math
Far Below	below 471	below 25
Approaches	471	25
Meets	487	31
Exceeds	537	52

Table 5: Cambridge IGCSE Mathematics Equating Table

Camb	AIMS	Camb	AIMS	Camb	AIMS
20	442	45	518	70	621
21	453	46	521	71	627
22	459	47	524	72	633
23	464	48	526	73	639
24	469	49	529	74	644
25	472	50	532	75	650
26	475	51	533	76	656
27	478	52	538	77	661
28	480	53	539	78	665
29	483	54	545	79	669
30	485	55	549	80	672
31	487	56	552	81	675
32	488	57	556	82	677
33	492	58	560	83	678
34	494	59	565	84	679
35	496	60	569	85	680
36	498	61	573	86	680
37	500	62	578	87	680
38	502	63	582	88	681
39	504	64	587	89	681
40	506	65	592	90	681
41	509	66	598	91	681
42	511	67	603	92	681
43	513	68	609	Approach	Meets
44	516	69	615	Exceeds	

Linking Arizona AIMS HS Reading Scale Scores with Cambridge IGCSE English First Language PUM Scores

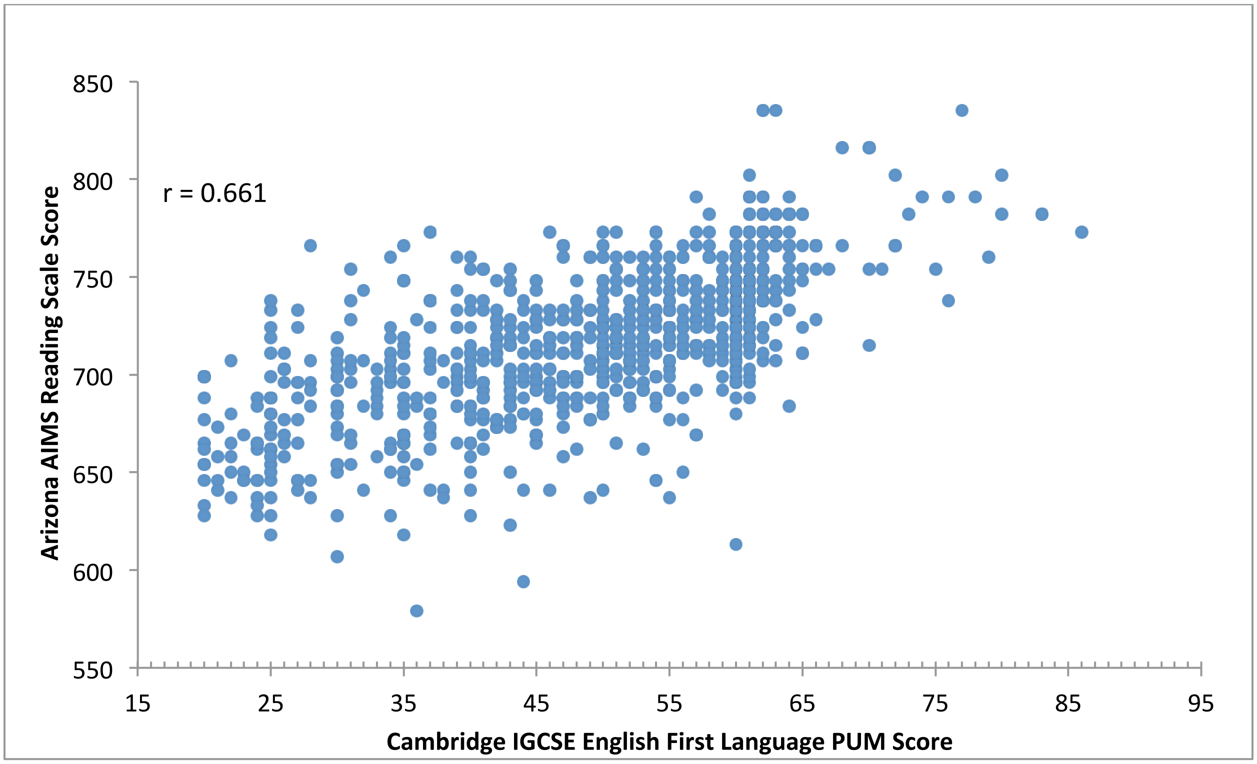
In 2012-2013 Arizona had 879 students who took both the AIMS HS Reading and the Cambridge IGCSE English First Language examinations. These common students formed the basis for common group equipercentile equating. The AIMS HS Reading assessment is reported on a scale that ranges from 500 to 900 and has four performance reporting categories corresponding to the NCLB categories of *Minimal, Basic, Proficient,* and *Advanced*. The four performance categories are presented in Table 6.

Table 6: Arizona Reading Performance Level Descriptors
(From http://www.azed.gov/standards-development-assessment/aims/performance/)

Performance Level	General Descriptor
Exceeds the Standard (773 - 900)	Students who score in this level illustrate a superior academic performance as evidenced by achievement that is substantially beyond the goal for all students. Students who perform at this level demonstrate strong analytical and inferential skills in comprehending more challenging and complex text. They are able to determine the meaning of vocabulary using minimal context clues, correctly utilize colloquialisms and historical jargon, and use knowledge of modes to interpret text.
Meets the Standard (674 - 772)	Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the reading standard. Students who perform at this level are able to identify and analyze literary elements such as theme, character, setting, plot, and point of view in complex texts. They will be able to support conclusions drawn from ideas and concepts in expository text and synthesize information from multiple sources to draw conclusions.
Approaches the Standard (627 – 673)	Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show a basic understanding of literary elements, making relevant inferences, and interpreting graphic sources of information to support ideas. They can compare (and contrast) classic works of literature that deal with similar topics and problems. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.
Falls Far Below the Standard (500 - 626)	Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's reading standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

A scatter plot of the AIMS HS Reading scores versus the Cambridge IGCSE English First Language scores is given in Figure 3. The two scores show a moderate association, with a correlation of 0.661.

Figure 3: Scatterplot of AIMS HS Reading Scale Score vs. Cambridge IGCSE English First Language PUM Score



Our first step in this linking was to examine the distributions and smooth the AIMS HS Reading and Cambridge IGCSE English First Language distributions. Pre-smoothing of the Cambridge IGCSE English First Language and AIMS HS Reading distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 7.

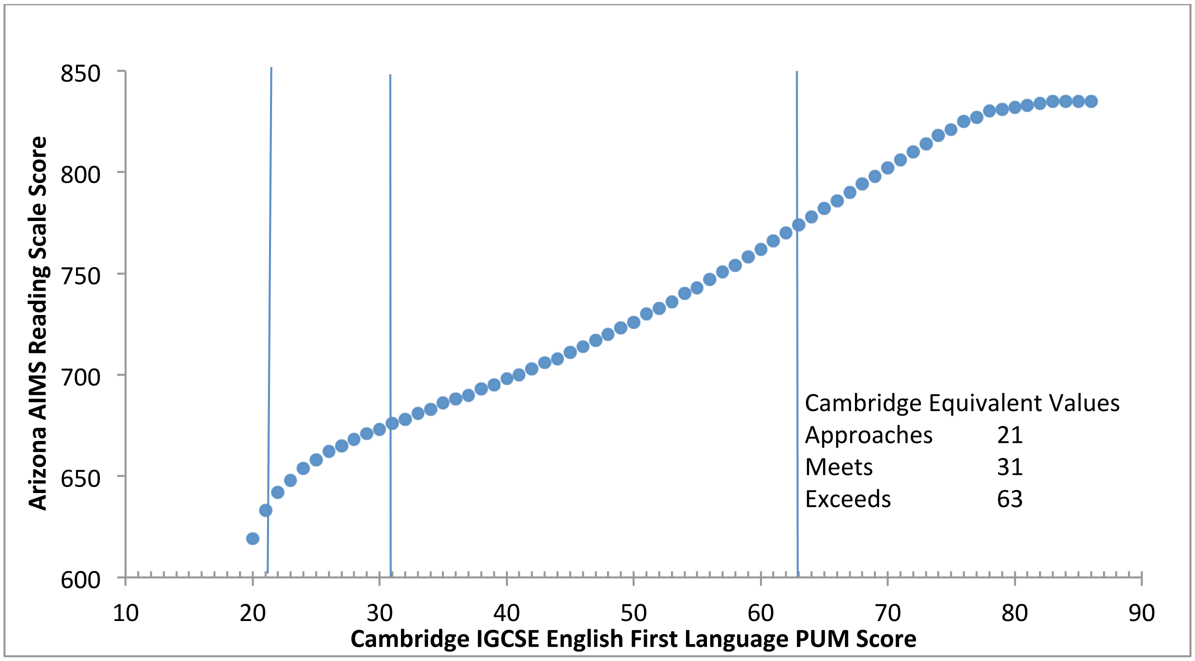
From our analysis we determined that a polynomial log-linear model of order 8 had the best fit for the AIMS HS Reading distribution and the Cambridge IGCSE First Language distribution.

Figure 4 presents the equipercentile linking relationship between the Arizona AIMS HS Reading examination and the Cambridge IGCSE English First Language examination. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 4. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Table 7: Fit for Reading/English Pre-smoothing

	AIMS Reading	Camb English
Model	AIC	AIC
Log-lin C=2	688.66	258.37
Log-lin C=3	589.32	251.73
Log-lin C=4	591.31	251.99
Log-lin C=5	546.82	249.02
Log-lin C=6	541.33	247.87
Log-lin C=7	533.68	249.86
Log-lin C=8	513.31	247.61

Figure 4: Cambridge IGCSE English First Language and AIMS HS Reading Equipercentile Equating Relationship



Graphically, Figure 4 shows the Cambridge IGCSE English First Language equivalent approaches, meets, or exceeds standards thresholds for the Arizona AIMS HS Reading exam. Table 8 summarizes the reporting thresholds, while Table 9 displays the full presentation of Cambridge IGCSE English First Language equivalent scores. As before, the Cambridge values for the “Approaches”, “Meets” or “Exceeds” thresholds are the first Cambridge equivalent values with a Cambridge equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Table 8: Cambridge Equivalent Thresholds

Threshold	AIMS Reading	Camb English
Far Below	below 627	below 21
Approaches	627	21
Meets	674	31
Exceeds	773	63

Table 9: Cambridge IGCSE Reading Equating Table

Camb	AIMS	Camb	AIMS	Camb	AIMS
20	619	43	706	66	786
21	631	44	708	67	790
22	642	45	711	68	794
23	648	46	714	69	798
24	654	47	717	70	802
25	658	48	720	71	806
26	662	49	723	72	810
27	665	50	726	73	814
28	668	51	730	74	818
29	671	52	733	75	821
30	673	53	736	76	825
31	676	54	740	77	827
32	677	55	743	78	830
33	681	56	747	79	831
34	683	57	751	80	832
35	686	58	754	81	833
36	688	59	758	82	834
37	690	60	762	83	835
38	693	61	766	84	835
39	695	62	770	85	835
40	698	63	773	86	835
41	700	64	778	Approach	Meets
42	703	65	782	Exceeds	

Linking Arizona AIMS HS Writing Scale Scores with Cambridge IGCSE English First Language PUM Scores

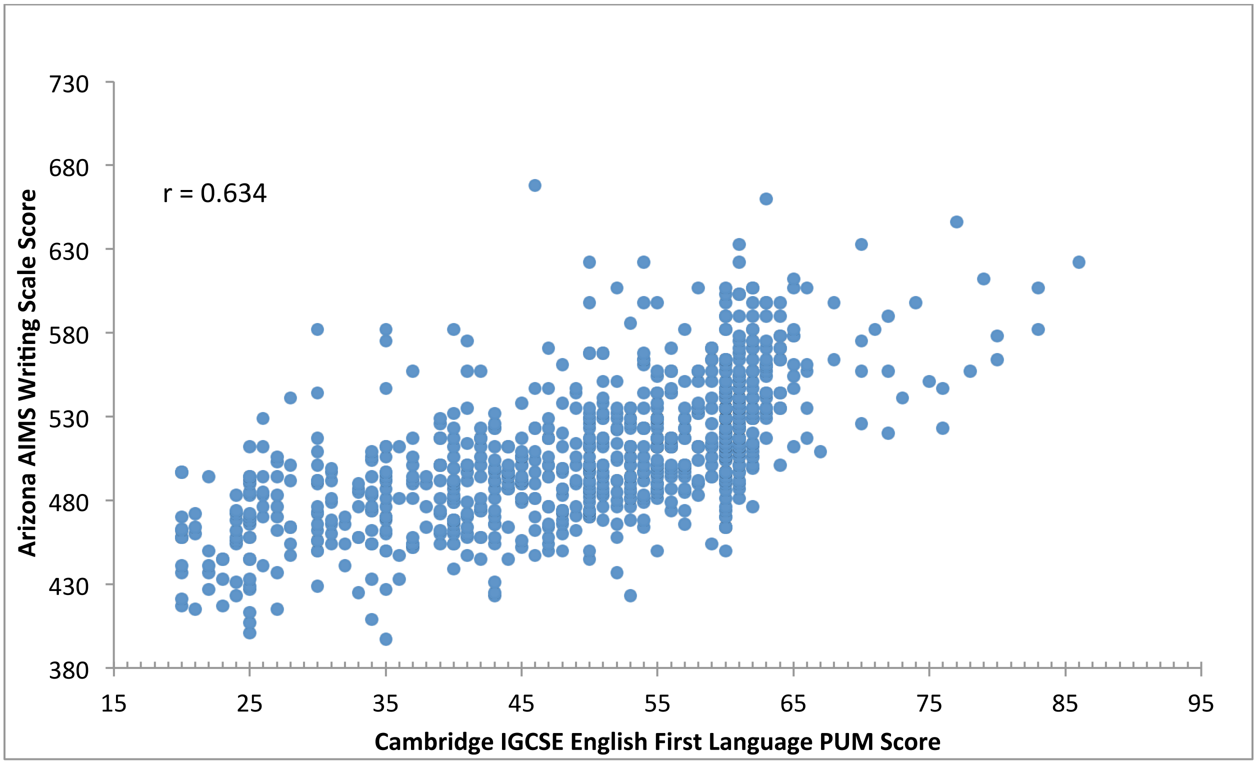
In 2012-2013 Arizona had 879 students who took both the AIMS HS Writing and the Cambridge IGCSE English First Language assessment. These common students formed the basis for common group equipercentile equating. The AIMS HS Writing assessment is reported on a scale that ranges from 300 to 700 and has four performance reporting categories corresponding to the NCLB categories of *Minimal, Basic, Proficient,* and *Advanced*. The four performance categories are presented in Table 10.

Table 10: Arizona Writing Performance Level Descriptors
(From http://www.azed.gov/standards-development-assessment/aims/performance/)

Performance Level	General Descriptor
Exceeds the Standard (587 - 700)	Students who score at this level show skillful performance in written communication as evidenced by performing substantially beyond the achievement goal for all students. Students who perform at this level consistently demonstrate the ability to identify and apply superior written communications by exhibiting a strong command of language including: clear, controlled ideas and organization, wide sentence variety, and impressive control of voice, conventions, and word choice.
Meets the Standard (480 - 586)	Students who score at this level show appropriate and acceptable performance at the Grade 10 Writing Standard. Students who perform at this level frequently demonstrate the ability to identify and apply adequate written communication by exhibiting a basic command of language including: clear ideas and organization, average sentence variety and functional control of voice, conventions, and word choice.
Approaches the Standard (433 - 479)	Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show a basic understanding of literary elements, making relevant inferences, and interpreting graphic sources of information to support ideas. They can compare (and contrast) classic works of literature that deal with similar topics and problems. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.
Falls Far Below the Standard (300 - 432)	Students who score at this level show less than adequate performance in written communication that falls significantly below the Grade 10 Writing Standard. Students who perform at this level unsatisfactorily demonstrate the ability to identify and apply basic written communications by exhibiting a weak command of language including: over simplistic or unclear ideas and organization, uncontrolled sentence variety, and highly limited control of voice, conventions, and word choice.

A scatter plot of the AIMS HS Writing scores versus the Cambridge IGCSE English First Language scores is given in Figure 5. The two scores show a moderate association, with a correlation of 0.634.

Figure 5: Scatterplot of AIMS HS Writing Scale Score vs. Cambridge IGCSE English First Language PUM Score



Our first step in this linking was to examine the distributions and smooth the AIMS HS Writing and Cambridge IGCSE English First Language distributions. Pre-smoothing of the Cambridge IGCSE English First Language and AIMS 10th grade Writing distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 11.

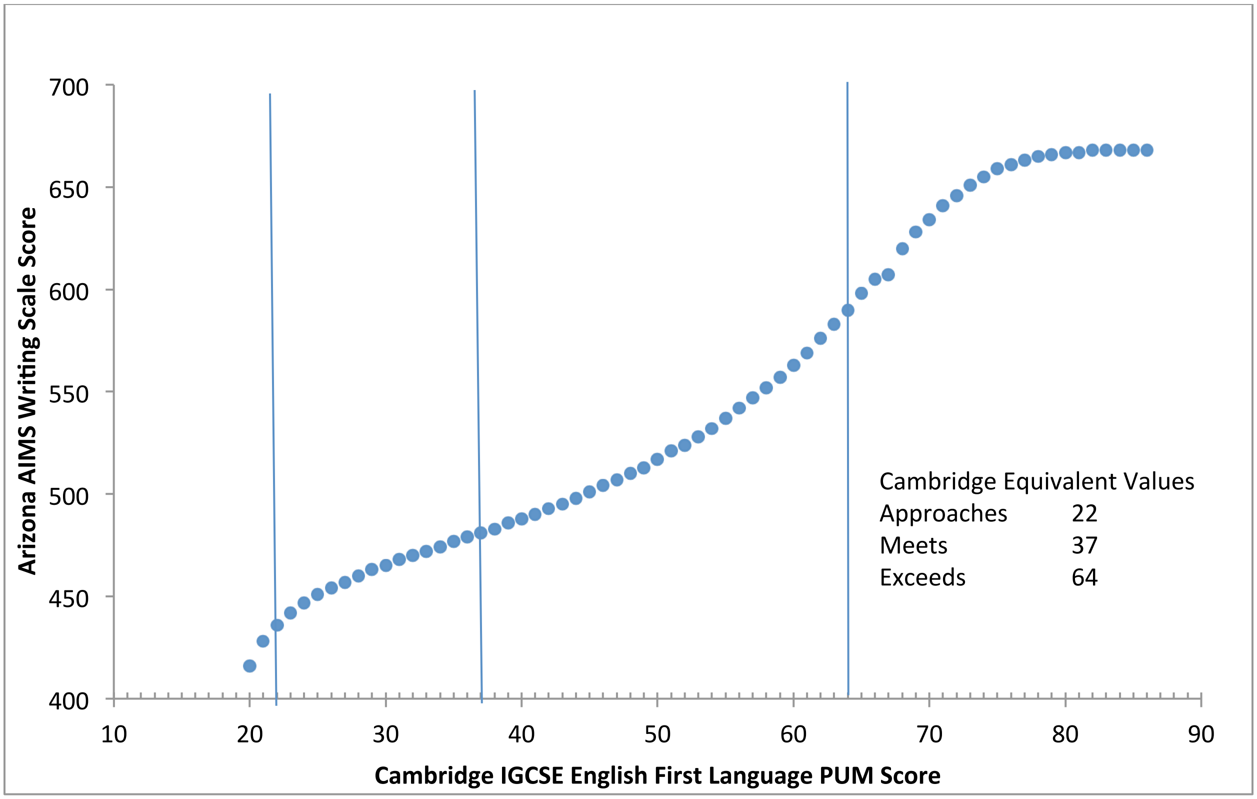
From our analysis we determined that a polynomial log-linear model of order 8 had the best fit for the AIMS HS Writing distribution, while a polynomial of order 8 produced the best fit for the Cambridge IGCSE English First Language distribution.

Figure 6 presents the equipercentile linking relationship between the Arizona AIMS HS Writing examination and the Cambridge IGCSE English First Language examination. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 6. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Table 11: Fit for Writing/English Pre-smoothing

	AIMS Writing	Camb English
Model	AIC	AIC
Log-lin C=2	745.40	258.37
Log-lin C=3	740.16	251.73
Log-lin C=4	738.15	251.99
Log-lin C=5	737.18	249.02
Log-lin C=6	736.95	247.87
Log-lin C=7	737.27	249.86
Log-lin C=8	734.16	247.61

Figure 6: Cambridge IGCSE English First Language and AIMS HS Writing Equipercentile Equating Relationship



Graphically, Figure 6 shows the Cambridge IGCSE English First Language equivalent approaches, meets, or exceeds standards thresholds for the Arizona AIMS HS Writing exam. Table 12 summarizes the reporting thresholds, while Table 13 displays the full presentation of Cambridge IGCSE English First Language equivalent scores. Again, the Cambridge values for the "Approaches", "Meets" or "Exceeds" thresholds are the first Cambridge equivalent values with a Cambridge equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Table 12: Cambridge Equivalent Thresholds

Threshold	AIMS Writing	Camb English
Far Below	below 433	below 22
Approaches	433	22
Meets	480	37
Exceeds	587	64

Table 13: Writing/English Equating Table

Camb	AIMS	Camb	AIMS	Camb	AIMS
20	416	43	495	66	605
21	428	44	498	67	607
22	433	45	501	68	620
23	442	46	504	69	628
24	447	47	507	70	634
25	451	48	510	71	641
26	454	49	513	72	646
27	457	50	517	73	651
28	460	51	521	74	655
29	463	52	524	75	659
30	465	53	528	76	661
31	468	54	532	77	663
32	470	55	537	78	665
33	472	56	542	79	666
34	474	57	547	80	667
35	477	58	552	81	667
36	479	59	557	82	668
37	480	60	563	83	668
38	483	61	569	84	668
39	486	62	576	85	668
40	488	63	583	86	668
41	490	64	590	Approach	Meets
42	493	65	598	Exceeds	

Linking Arizona AIMS HS Mathematics Scale Scores with ACT *QualityCore* Algebra I Scores and/or ACT *QualityCore* Geometry Scores

In 2012-2013 Arizona had 286 students who completed the AIMS HS Mathematics assessment and the ACT *QualityCore* Geometry and Algebra I assessments. These common students formed the basis for common group equipercentile equating. All ACT *QualityCore* assessments are reported on a scale that ranges from 125 to175. The question is which of the two assessments, ACT *QualityCore* Geometry or Algebra I, provide the best means for linking to the AIMS HS Mathematics assessment, or does a composite measure combining the Geometry and Algebra I scale scores produce a more desirable means for linking because the AIMS HS Mathematical assessment includes both algebra and geometry. Therefore, we present results for both ACT *QualityCore* Geometry and Algebra I scores and for a Composite scale consisting of the sum of the Geometry and Algebra I scores. Scatter plots of the AIMS HS Mathematics scores versus the ACT *QualityCore* Geometry and Algebra I scores are given in Figures 7 and 8. ACT *QualityCore* Geometry and Algebra I show a moderate correlation with AIMS HS Mathematics – 0.641 and 0.653 respectively. Figure 9 presents a scatterplot of the Composite scores versus the AIMS Mathematics scores. Not surprisingly, the Composite shows a stronger correlation (0.714) than the Geometry and Algebra I scores separately.

Figure 7: Scatterplot of AIMS HS Mathematics Scale Score vs. ACT *QualityCore* Geometry Scale Score

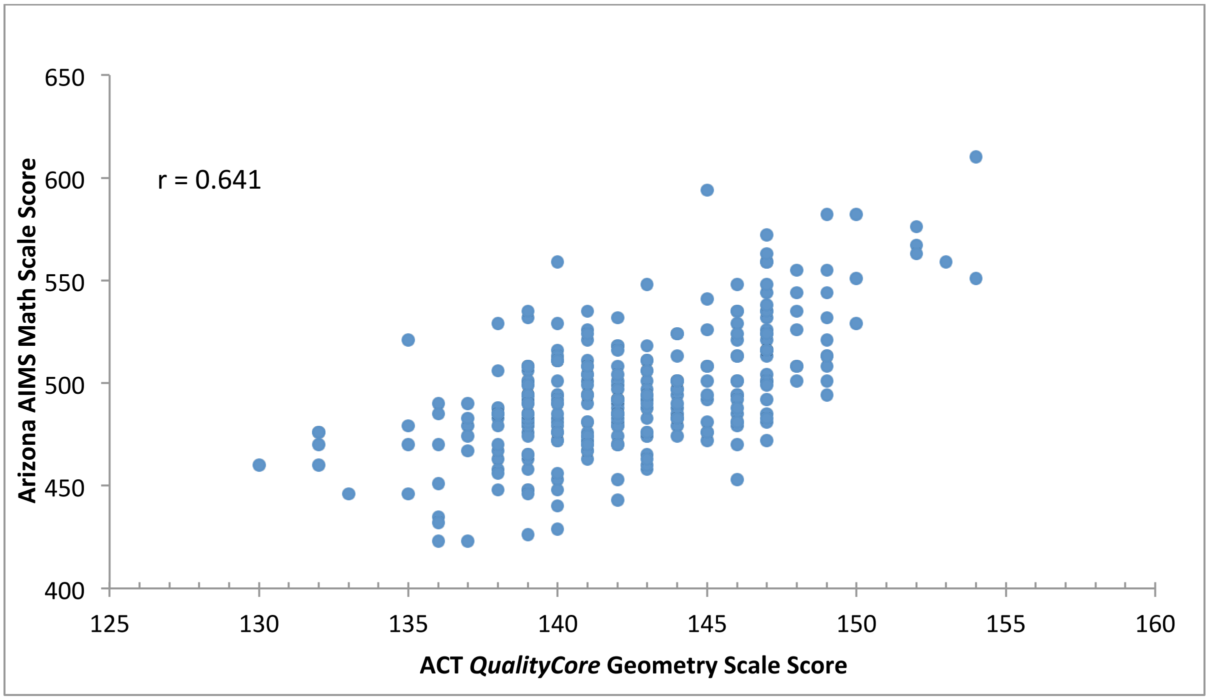


Figure 8: Scatterplot of AIMS HS Mathematics Scale Score vs. ACT *QualityCore* Algebra I Scale Score

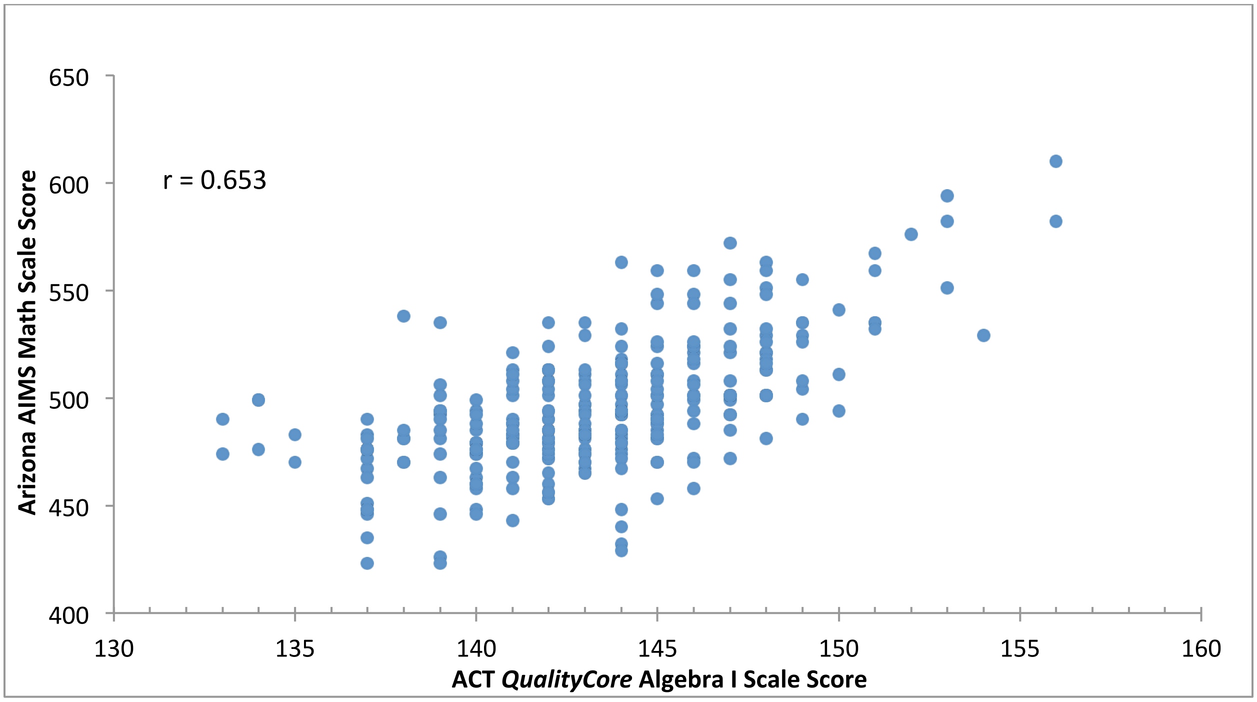
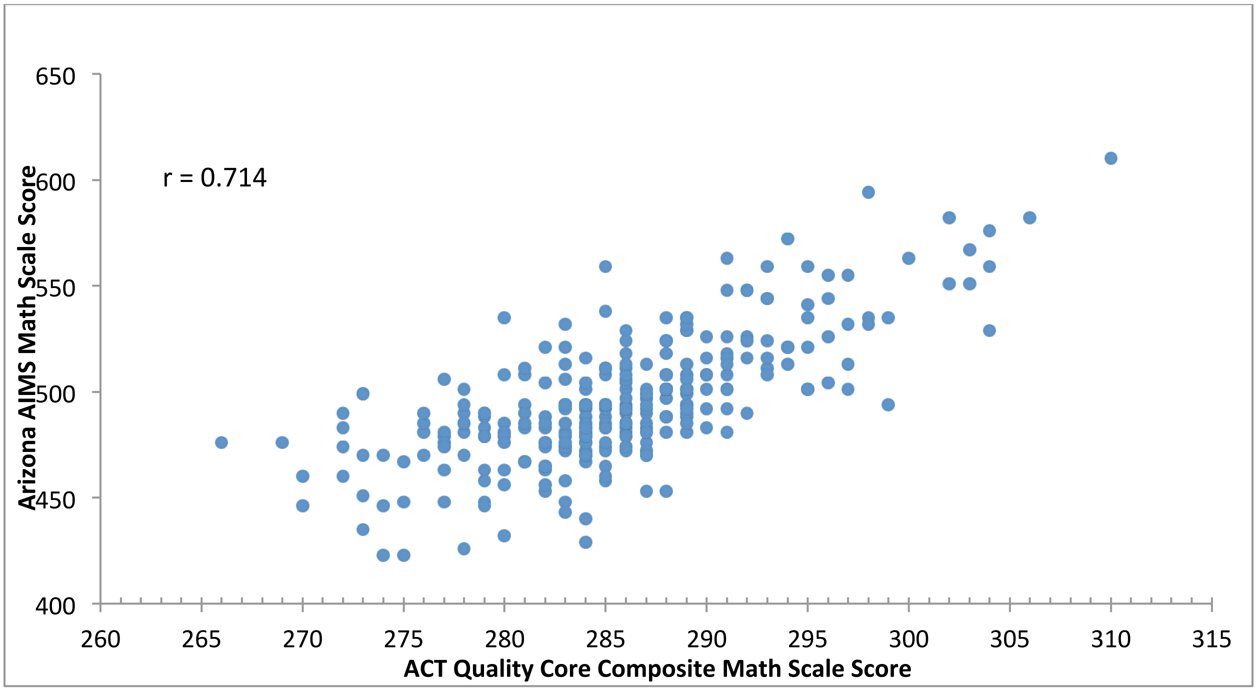


Figure 9: Scatterplot of AIMS HS Mathematics Scale Score vs. ACT *QualityCore* Composite Scale Score



Our first step in linking was to examine and smooth the AIMS HS Mathematics and ACT Geometry distributions. Pre-smoothing for both distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 14.

Table 14: Fit for Math Pre-smoothing

	AIMS Math	QC Geometry
Model	AIC	AIC
Log-lin C=2	264.15	140.63
Log-lin C=3	251.65	141.98
Log-lin C=4	251.89	142.45
Log-lin C=5	244.74	141.04
Log-lin C=6	244.45	141.62
Log-lin C=7	246.19	132.65
Log-lin C=8	247.40	129.99

From our analysis we determined that a polynomial log-linear model of order 6 had the best fit for the AIMS HS Mathematics exam distribution, while a polynomial of order 8 produced the best fit for the ACT Geometry math exam distribution.

Figure 10 presents the equipercentile linking relationship between the Arizona AIMS HS Mathematics examination and the ACT *QualityCore* Geometry. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 10. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Graphically, Figure 10 shows the ACT Geometry math equivalent “approaches”, “meets” and “exceeds” standards thresholds for the Arizona AIMS HS Mathematics examination. Table 15 summarizes the reporting thresholds, while Table 16 displays the full presentation of ACT Geometry math equivalent scores. Again, the ACT values for the “Approaches”, “Meets” or “Exceeds” thresholds are the first ACT equivalent values with an ACT equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Figure 10: ACT *QualityCore* Geometry and AIMS HS Mathematics Equipercentile Equating Relationship

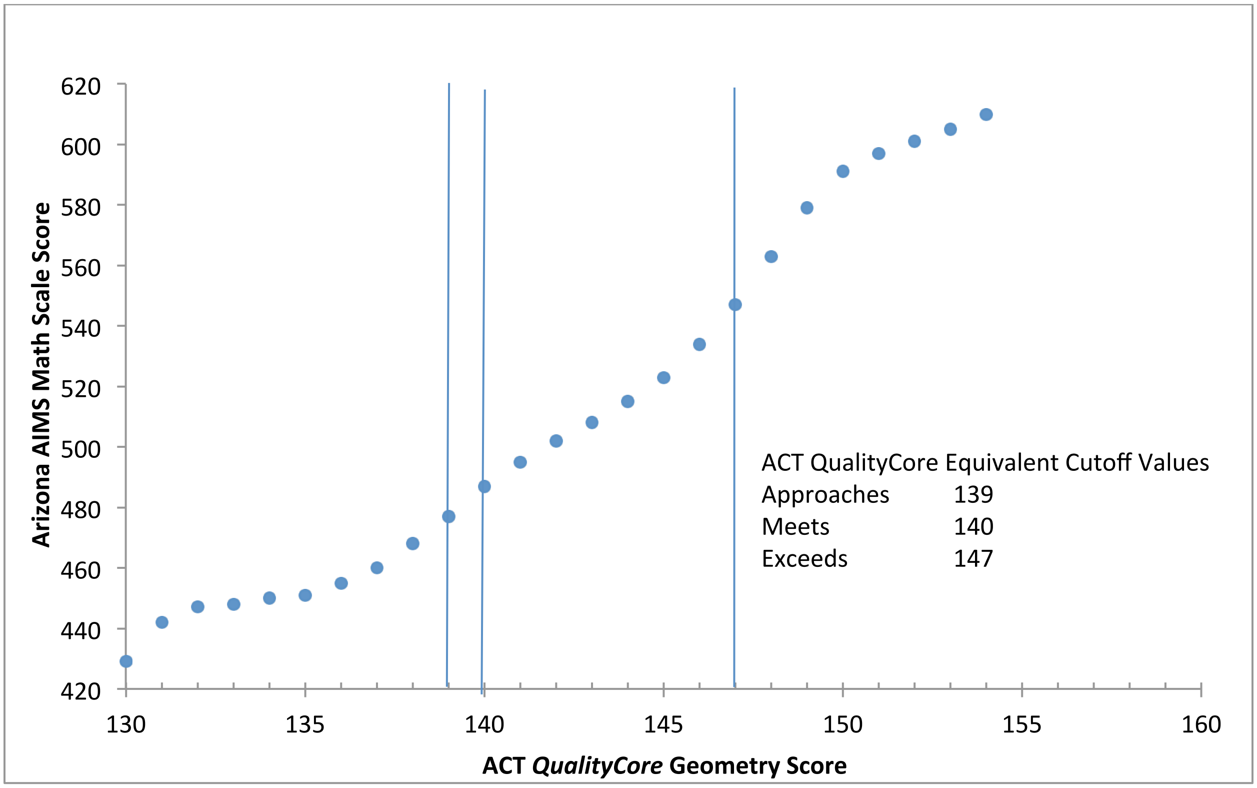


Table 15: ACT QC Geometry Equivalent Thresholds

Threshold	AIMS Math	ACT Geometry
Far Below	below 471	below 139
Approaches	471	139
Meets	487	140
Exceeds	537	147

Table 16: ACT *QualityCore* Geometry Equating Table

ACT Geom	AIMS	ACT Geom	AIMS	ACT Geom	AIMS
130	429	139	472	148	563
131	442	140	487	149	579
132	447	141	488	150	591
133	448	142	502	151	597
134	450	143	508	152	601
135	451	144	515	153	605
136	455	145	523	154	610
137	460	146	534	Approach	Meets
138	468	147	539	Exceeds	

Our next step in linking was to examine and smooth the AIMS HS Mathematics and ACT Algebra I distributions. Pre-smoothing for both distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 17.

Table 17: Fit for Math Pre-smoothing

	AIMS Math	QC Algebra
Model	AIC	AIC
Log-lin C=2	264.15	122.20
Log-lin C=3	251.65	123.48
Log-lin C=4	251.89	115.86
Log-lin C=5	244.74	117.54
Log-lin C=6	244.45	118.89
Log-lin C=7	246.19	120.88
Log-lin C=8	247.40	121.69

From our analysis we determined that a polynomial log-linear model of order 6 had the best fit for the AIMS HS Mathematics exam distribution, while a polynomial of order 4 produced the best fit for the ACT Composite math exam distribution.

Figure 11 presents the equipercentile linking relationship between the Arizona AIMS HS Mathematics examination and ACT *QualityCore* Algebra I. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 11. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Graphically, Figure 11 shows the ACT Algebra I math equivalent “approaches”, “meets” and “exceeds” standards thresholds for the Arizona AIMS HS Mathematics examination. Table 18 summarizes the reporting thresholds, while Table 19 displays the full presentation of ACT composite math equivalent scores. Again, the ACT values for the “Approaches”, “Meets” or “Exceeds” thresholds are the first ACT equivalent values with an ACT equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Figure 11: ACT *QualityCore* Algebra I and AIMS HS Mathematics Equipercentile Equating Relationship

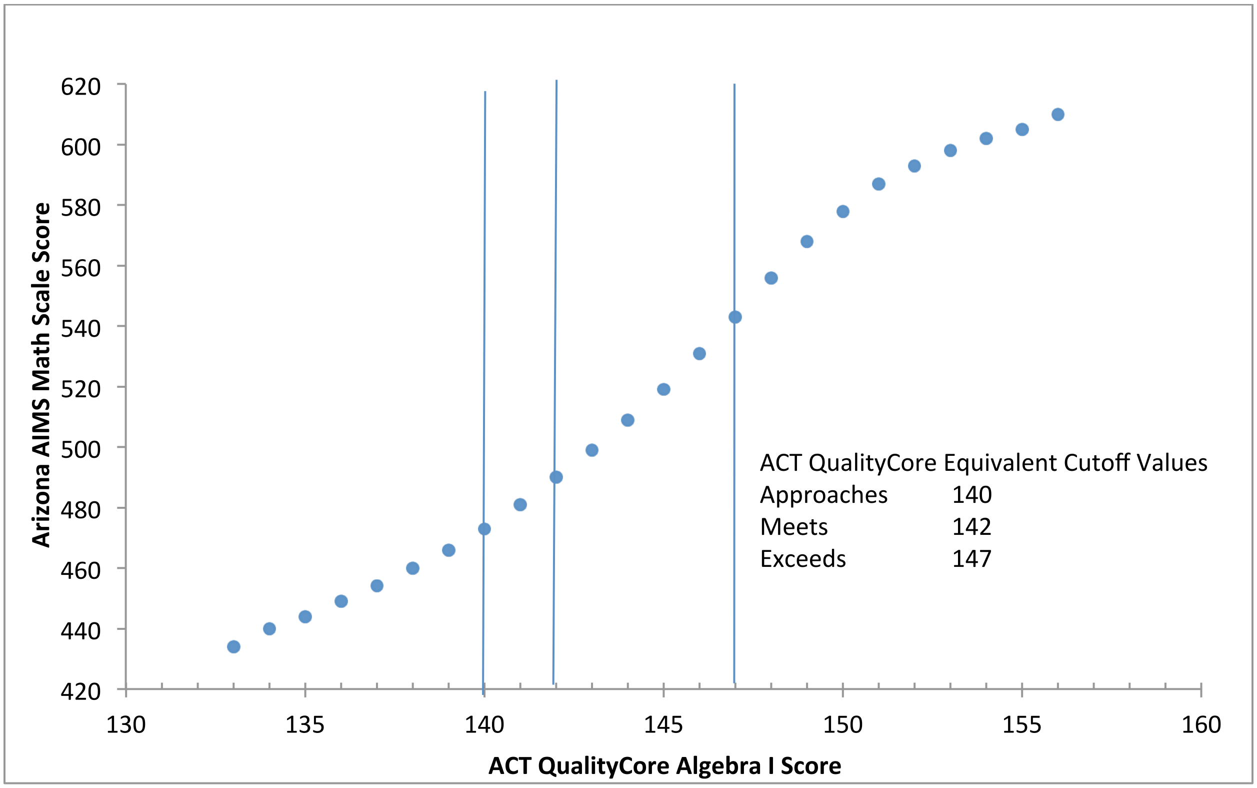


Table 18: ACT *QC Algebra* Equivalent Thresholds

Threshold	AIMS Math	ACT Algebra
Far Below	below 471	below 140
Approaches	471	140
Meets	487	142
Exceeds	537	147

Table 19: ACT *QualityCore* Algebra I Equating Table

ACT Alg	AIMS	ACT Alg	AIMS	ACT Alg	AIMS
133	434	142	488	151	587
134	440	143	499	152	593
135	444	144	509	153	598
136	449	145	519	154	602
137	454	146	531	155	605
138	460	147	539	156	610
139	466	148	556		
140	472	149	568	Approach	Meets
141	481	150	578	Exceeds	

Our next step in this analysis was to examine the distributions and smooth the AIMS Mathematics and ACT *QualityCore* Composite Mathematics distribution. Pre-smoothing of the *QualityCore* Composite Mathematics and the AIMS 10th grade Mathematics distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 20.

Table 20: Fit for Math Pre-smoothing

	AIMS Math	QC Math Comp
Model	AIC	AIC
Log-lin C=2	264.15	178.46
Log-lin C=3	251.65	175.48
Log-lin C=4	251.89	165.34
Log-lin C=5	244.74	165.49
Log-lin C=6	244.45	161.77
Log-lin C=7	246.19	163.19
Log-lin C=8	247.40	163.67

From our analysis we determined that a polynomial log-linear model of order 6 had the best fit for the AIMS Mathematics exam distribution, while a polynomial of order 6 produced the best fit for the ACT *QualityCore* Composite Mathematics exam distribution.

Figure 12 presents the equipercentile linking relationship between the Arizona AIMS Mathematics scores and the ACT *QualityCore* Composite Mathematics scores. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 12. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Graphically, Figure 12 shows the ACT *QualityCore* Composite Mathematics equivalent approaches, meets or exceeds standards thresholds for the Arizona AIMS Mathematics examination. Table 21 summarizes the reporting thresholds, while Table 22 displays the full presentation of ACT Composite Mathematics equivalent scores. Again, the ACT values for the "Approaches", "Meets" or "Exceeds" thresholds are the first ACT equivalent values with an ACT equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Figure 12: ACT *QualityCore* Composite and AIMS HS Mathematics Equipercentile Equating Relationship

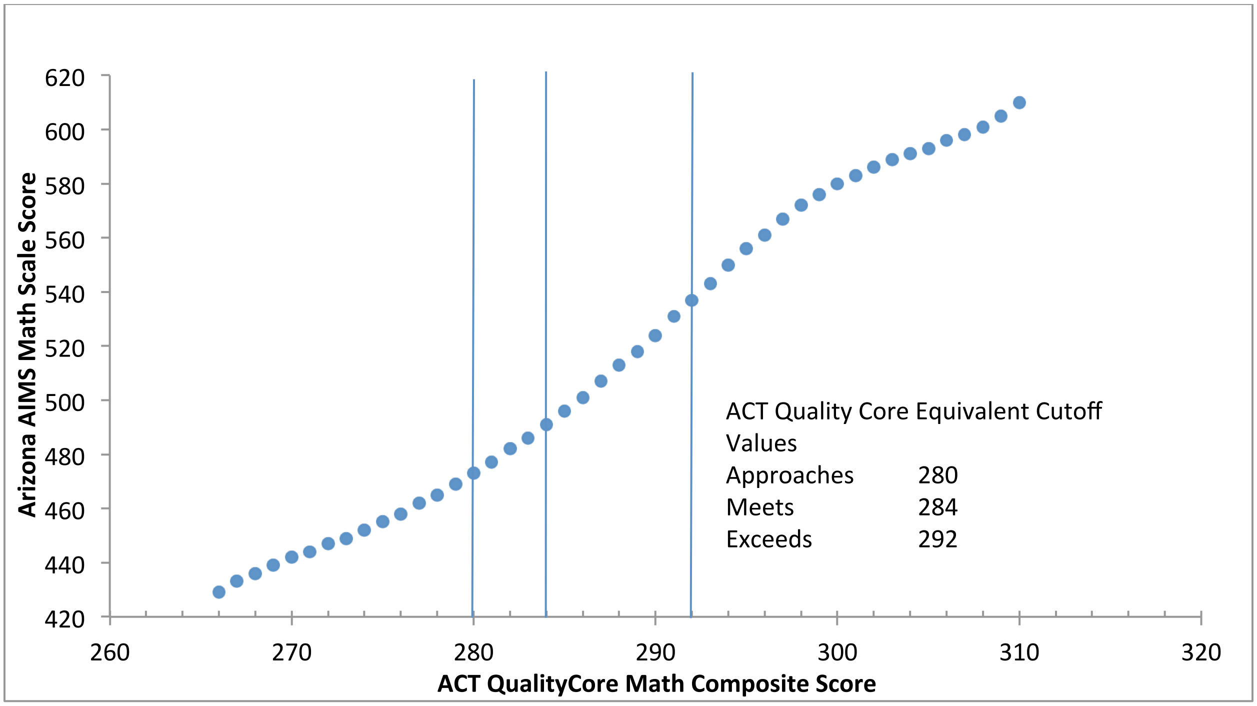


Table 21: ACT *QC* Comp Equivalent Thresholds

Threshold	AIMS Math	ACT Math Comp
Far Below	below 471	below 280
Approaches	471	280
Meets	487	284
Exceeds	537	292

Table 22: ACT *QualityCore* Composite Equating Table

ACT Math	AIMS	ACT Math	AIMS	ACT Math	AIMS
266	429	281	477	296	561
267	433	282	482	297	567
268	436	283	486	298	572
269	439	284	491	299	576
270	442	285	496	300	580
271	444	286	501	301	583
272	447	287	507	302	586
273	449	288	513	303	589
274	452	289	518	304	591
275	455	290	524	305	593
276	458	291	531	306	596
277	462	292	537	307	598
278	465	293	543	308	601
279	469	294	550	309	605
280	473	295	556	310	610
Approach	Meets	Exceeds			

We were able to successfully link the ACT *QualityCore* Algebra I, Geometry and Composite scores to the AIMS HS Mathematics examination. Consequently, Arizona policy makers have several choices moving forward: (1) use only the ACT *QualityCore* Algebra I exam; (2) use only the ACT *QualityCore* Geometry exam; (3) use both examinations; (4) adopt an either/or scenario using both exams; or (5) use the Composite score consisting of the sum of the Algebra I and Geometry scores. Since all three scores were successfully linked with the AIMS HS Mathematics test this is ultimately a policy decision that rests with the Arizona State Board of Education with each option offering different benefits.

Linking Arizona AIMS HS Reading Scale Scores with ACT *QualityCore* English Scale Scores

In 2012-2013 Arizona had 340 students who took both the AIMS HS Reading and the ACT *QualityCore* English 10 assessments. These common students formed the basis for common group equipercentile equating. A scatter plot of the AIMS HS Reading scores versus the ACT *QualityCore* English 10 scores is given in Figure 13. The two scores show a moderately strong association, with a correlation of 0.708.

Our first step in this linking was to examine the distributions and smooth the AIMS HS Reading and ACT *QualityCore* English 10 distributions. Pre-smoothing of the ACT *QualityCore* English 10 and AIMS HS Reading distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 23.

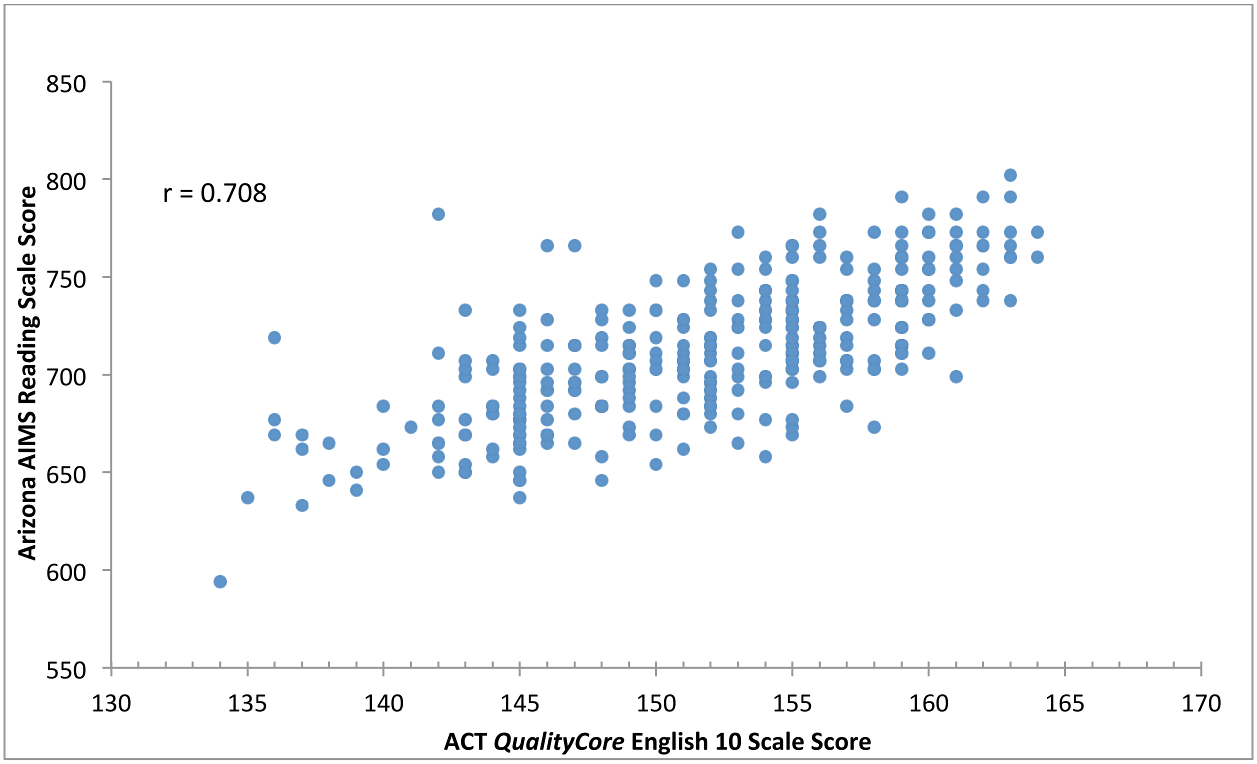
From our analysis we determined that a polynomial log-linear model of order 2 had the best fit for the AIMS HS Reading distribution, while a polynomial of order 2 produced the best fit for the ACT *QualityCore* English 10 distribution.

Table 23: Fit for Reading/English Pre-smoothing

	AIMS Reading	ACT QC English
Model	AIC	AIC
Log-lin C=2	181.95	187.92
Log-lin C=3	183.09	188.81
Log-lin C=4	184.92	189.84
Log-lin C=5	184.44	189.25
Log-lin C=6	186.38	190.92
Log-lin C=7	187.64	192.85
Log-lin C=8	188.32	188.62

Figure 14 presents the equipercentile linking relationship between the Arizona AIMS HS Reading examination and the ACT *QualityCore* English 10 examination. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 14. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Figure 13: Scatterplot of AIMS HS Reading Scale Score vs. ACT *QualityCore* English 10 Scale Score



Graphically, Figure 14 shows the ACT *QualityCore* English 10 equivalent “approaches”, “meets” and “exceeds” standards thresholds for the Arizona AIMS HS Reading exam. Table 24 summarizes the reporting thresholds, while Table 25 displays the full presentation of ACT *QualityCore* English 10 equivalent scores. Again, the ACT values for the “Approaches”, “Meets” or “Exceeds” thresholds are the first ACT equivalent values with an ACT equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Table 24: ACT *QualityCore* Equivalent Thresholds

Threshold	AIMS Reading	ACT QC English
Far Below	below 627	below 137
Approaches	627	137
Meets	674	144
Exceeds	773	161

Figure 14: ACT *QualityCore* English 10 and AIMS HS Reading Equipercentile Equating Relationship

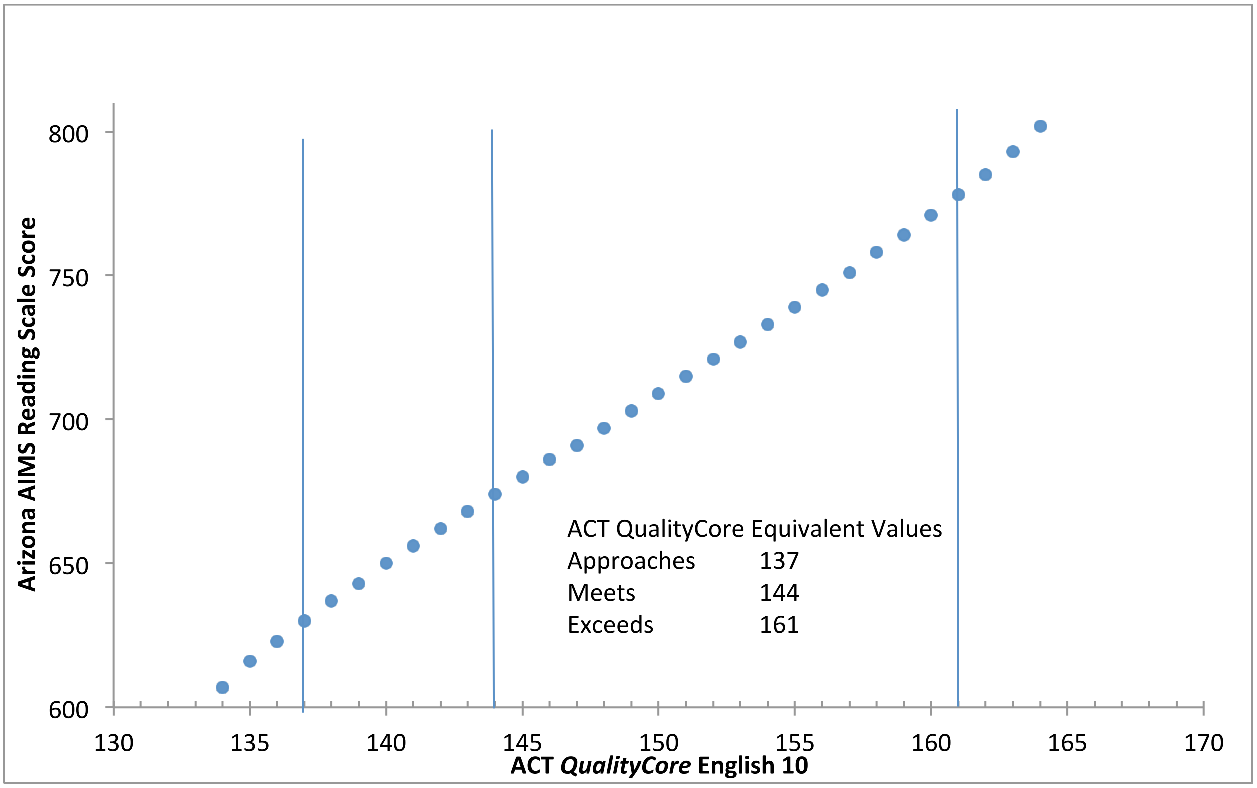


Table 25: Reading/English Equating Table

ACT QC	AIMS	ACT QC	AIMS	ACT QC	AIMS
134	607	145	677	156	745
135	616	146	686	157	751
136	623	147	691	158	758
137	630	148	697	159	764
138	631	149	703	160	771
139	643	150	709	161	773
140	650	151	715	162	785
141	656	152	721	163	793
142	662	153	727	164	802
143	668	154	733	Approach	Meets
144	674	155	739	Exceeds	

Linking Arizona AIMS HS Writing Scale Scores with ACT *QualityCore* English 10 Scale Scores

In 2012-2013 Arizona had 340 students who took both the AIMS HS Writing and the ACT *QualityCore* English 10 assessment. These common students formed the basis for common group equipercentile equating. A scatter plot of the AIMS HS Writing scores versus the ACT *QualityCore* English 10 scores is given in Figure 15. The two scores show a moderate association, with a correlation of 0.675 (see below).

Our first step in this linking was to examine the distributions and smooth the AIMS HS Writing and ACT *QualityCore* English distributions. Pre-smoothing of the ACT *QualityCore* English and AIMS HS Writing distributions was done by polynomial log-linear models. Models of various orders were tested to determine which produced the best fit and the distributional plots were then examined for indications of over-fitting. Polynomial models of orders from 2 to 8 were fit to each distribution. The Akaike Information Criterion (AIC) fit statistics are given in Table 26.

From our analysis we determined that a polynomial log-linear model of order 4 had the best fit for the AIMS HS Writing distribution, while a polynomial of order 2 produced the best fit for the ACT *QualityCore* English 10 distribution.

Figure 15: Scatterplot of AIMS HS Writing Scale Score vs. ACT *QualityCore* English 10 Scale Score

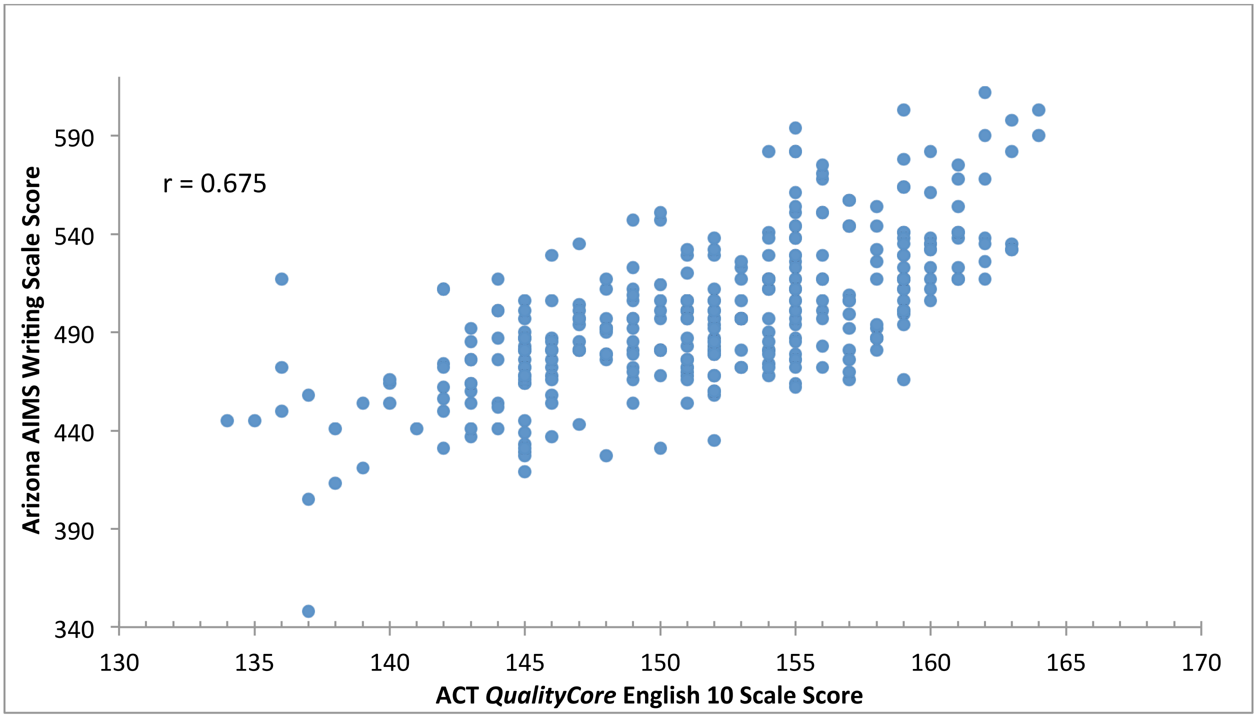
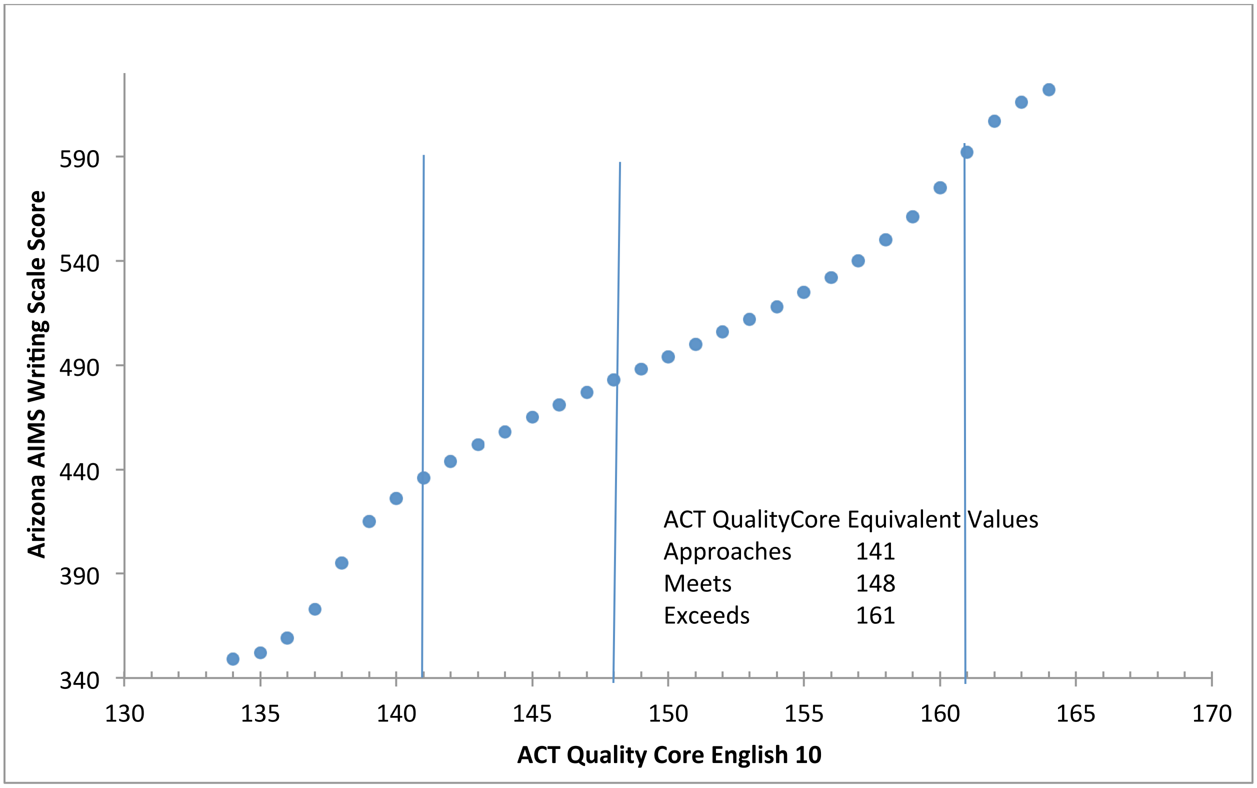


Table 26: Fit for Writing/English Pre-smoothing

	AIMS Writing	ACT QC English
Model	AIC	AIC
Log-lin C=2	398.18	187.92
Log-lin C=3	397.03	188.81
Log-lin C=4	376.10	189.84
Log-lin C=5	378.09	189.25
Log-lin C=6	379.23	190.92
Log-lin C=7	380.10	192.85
Log-lin C=8	382.10	188.62

Figure 16 presents the equipercentile linking relationship between the Arizona AIMS HS Writing examination and the ACT *QualityCore* English 10 examination. We originally proposed to conduct post-smoothing using kernel density estimation, but the equipercentile plot of the smoothed scores revealed that simple linear interpolation produced a sufficiently smooth relationship as seen in Figure 16. Any lack of smoothness in this plot is a result of the discrete nature of the score reporting for the respective tests.

Figure 16: ACT *QualityCore* English 10 and AIMS HS Writing Equipercentile Equating Relationship



Graphically, Figure 16 shows the ACT *QualityCore* English 10 equivalent approaches, meets, or exceeds standards thresholds for the Arizona AIMS HS Writing exam. Table 27 summarizes the reporting thresholds, while Table 28 displays the full presentation of ACT *QualityCore* English 10 equivalent scores. Again, the ACT values for the "Approaches", "Meets" or "Exceeds" thresholds are the first ACT equivalent values with an ACT equivalent AIMS value that either equals or exceeds the AIMS threshold in question.

Table 27: ACT *QualityCore* Equivalent Thresholds

Threshold	AIMS Writing	ACT QC English
Far Below	below 433	below 141
Approaches	433	141
Meets	480	148
Exceeds	587	161

Table 28: Writing/English Equating Table

ACT QC	AIMS	ACT QC	AIMS	ACT QC	AIMS
134	349	145	465	156	532
135	352	146	471	157	540
136	359	147	477	158	550
137	373	148	480	159	561
138	395	149	488	160	575
139	415	150	494	161	588
140	426	151	500	162	607
141	433	152	506	163	616
142	444	153	512	164	622
143	452	154	518	Approach	Meets
144	458	155	525	Exceeds	

Conclusions/Recommendations Going Forward

Using the equating methods described above we were able to successfully link the Cambridge IGCSE examinations and ACT *QualityCore* Examinations with equivalent scores on the Arizona AIMS HS examinations. Moreover, for each examination we were also able to identify corresponding NCLB thresholds for the state ("Approaches," "Meets" and/or "Exceeds" standards). Going forward we intend to confirm and reanalyze these results with additional waves of data. Specifically we can cross-validate the estimated thresholds to determine if they are stable from year to year.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.). *Educational Measurement (2nd Edition).* Washington, DC: American Council on Education.

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W. & Hemphill, F.C. (1999). *Uncommon Measures: Equivalence and Linkage among Educational Tests.* Washington, D.C.: National Academy Press.

Hansen, B.B. & Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15(3), 609-627.*

Kolen, M.J. & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* New York, NY: Springer.

Livingston, S.A. & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46(3), 330-343.*

Livingston, S.A. & Kim, S. (2011). New approaches to equating in small samples. In A. A. von Davier (Ed.). *Statistical Models for Test Equating, Scaling, and Linking.* New York, NY: Springer.

von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004). *The Kernel Method of Test Equating.* New York, NY: Springer.